in step S504 (step S505).

Then, processor 205 determines whether or not the determined spectral frequency distortion coefficient corresponds to a middle value of the three coefficients (step S506).

When the spectral frequency distortion coefficient providing the most similar distance does not correspond to the middle value in step S506, processor 205 determines whether or not the coefficient providing the most similar distance corresponds to a maximum value of the three distortion coefficients (step S508). When the spectral frequency distortion coefficient corresponds to the maximum value, processor 205 adds $\Delta\alpha_2$ to all of the three distortion coefficients to be calculated (step S509), and returns to the spectral frequency distortion calculation of step S502. The $\Delta\alpha_2$ is preferably 0.001 to 0.1000 when a sampling frequency of speech sounds is 10kHz, and the present embodiment employs $\Delta\alpha_2 = 0.02$. When the spectral frequency distortion coefficient does not correspond to the maximum value, processor 205 subtracts 0.02 from all of the three distortion coefficients to be calculated (step S510), and returns to the spectral frequency distortion calculation of step S502. Step S502 through step S510 are repeated until it is determined that the spectral frequency distortion coefficient providing the most similar distance corresponds to the middle value of the three distortion coefficients in step S506. When the spectral frequency distortion coefficient corresponds to the middle value, speaker adaptation processor 205 determines that the spectral frequency distortion coefficient is an optimum distortion coefficient (step S507), and finishes its process.

Speech recognition unit 206 receives, from acoustic analysis unit 202, a second utterance part of the LPC cepstral coefficient vector obtained by acoustically analyzing the utterance by the user. Unit 206 also receives, from pattern selection unit 204, the trained pattern to be selected determined by pattern selection unit 204 and the recognition result of the vocabulary indicating the controlled device. Unit 206 further receives, from speaker adaptation processor 205, the spectral frequency distortion coefficient determined by processor 205. Unit 206 performs the distortion calculation of spectral frequency for the received second utterance part of the LPC cepstral coefficient vector using the spectral frequency distortion coefficient determined by processor 205. At this time, unit 206 finds a distance between an actual word and a device control word registered in word lexicon 208. In other words, unit 206 determines, using the simplified Mahalanobis' distance

equation, the distances between all device control words and actual words of the devices for which a distance for the second utterance by the user has been selected. When category 2 is selected and the LPC cepstral coefficient vectors for control words 1 through 30 for the optimum spectral frequency distortion coefficient are X_1 through X_{30}, distance L is determined as follows.

$$L_{5nk} = \sum_{i,time} B_{5k_i} - 2\vec{A}_{5k_i}^t \cdot \vec{X}_n$$

where;

$$\vec{A}_{5k} = \vec{W}^{-1} \cdot \vec{\mu}_{5k} - \vec{W}^{-1} \cdot \vec{\mu}_x$$

$$B_{5k} = \vec{\mu}_{5k}^t \cdot \vec{W}^{-1} \cdot \vec{\mu}_{5k} - \vec{\mu}_x^t \cdot \vec{W}^{-1} \cdot \vec{\mu}_x$$

$\vec{\mu}_{5k}$ is an average value of LPC cepstral coefficient vectors of state (k) (phoneme order or time sequences) for category 2,

$\vec{\mu}_x$ is an average value of LPC cepstral coefficient vectors of all utterances by training speakers in category 2,

$\vec{W}$ is a covariance value of LPC cepstral coefficient vectors of all utterances by the training speakers in category 2, and

$\vec{X}_n$ is an LPC cepstral coefficient vector provided by the spectral frequency distortion calculation for a device control word n, and n is 1 through 30.

In this equation, the distance is calculated assuming all utterances in the corresponding category to be the entire distribution. This calculation thus provides a more reliable distance than a calculation in which all utterance in various categories are assumed to be the entire distribution. The equation is therefore extremely effective for the recognition of a device control word.

Control signal output unit 207 receives from speech recognition unit 206 recognition results of the following vocabularies: one indicating the controlled device determined by pattern selection unit 204; and the other

indicating the control content determined by speech recognition unit 206. Output unit 207 supplies a signal indicating the control content to the controlled device.

The present invention uses one kind of device control word files 210 — part of word lexicon 608, but different kind of device control word files may be used for each controlled device.   In this case, a number of vocabularies used for comparing distances is limited, so that a recognition time can be reduced, and yet, recognition performance is improved.

The voice controller shown in Fig. 2 except for sound input unit 201 and control signal output unit 207 is hereinafter called speech recognition apparatus 211.

Table 1 shows speech recognition performance, in the case that only one device control word "Television" and 30 device control words such as "Channel one" and "Increase sound" are prepared.   The recognition performance is represented by speech recognition ratios.   This test was performed under a noise environment in which a signal-to-noise (S/N) ratio was 15dB, and a sampling frequency of speech sound was set at 10 kHz.

Table 1

Adaptation methods	Ages of users		
	12 or lower	13 through 64	65 or higher
No adaptation	78.7%	88.7%	82.3%
Adaptation by only pattern selection	84.4%	89.2%	84.5%
Adaptation by only vocal tract length normalization by distortion of spectral	86.8%	93.9%	83.9%
Adaptation by pattern selection and distortion of spectral frequency	90.0%	94.6%	87.5%

The adaptation methods in Table 1 are described.

In the case of "No adaptation", a single trained pattern is used, all training speakers are not categorized, and spectral frequency of a user's speech sound is not distorted.

In the case of "Adaptation by only pattern selection", trained patterns are generated in response to ages of speakers, and pattern by-characteristic selection unit 204 selects a trained pattern.   Spectral frequency of a user's